24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery

ELSEVIER

Journal of Clinical Epidemiology 57 (2004) 268–276

Journal of Clinical Epidemiology

24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery

Raymond W.J.G. Ostelo[a,*], Henrica C.W. de Vet[a], Dirk L. Knol[a], Piet A. van den Brandt[b]

[a]*EMGO Institute, VU University Medical Center, Van der Boechorststraat 7 1081 BT, Amsterdam, The Netherlands*
[b]*Department of Epidemiology, Maastricht University, P.O. Box 6.6, 6200 MD, Maastricht, The Netherlands*

Accepted 13 September 2003

Abstract

Objective: Measurement properties of questionnaires should be based on samples of populations on whom these measurements will be used. The purpose of this study is to establish an evidence based recommendation regarding the use of functional status questionnaires in patients following a lumbar disc surgery by a direct comparison of the reproducibility and responsiveness.

Study Design and Setting: The measurement properties of six functional status questionnaires were assessed: 1) Roland-Morris Disability Questionnaire (RDQ-24), 2) Modified Roland-Morris Disability Questionnaire (MRDQ), 3) short Roland-Morris Disability Questionnaire (RM-18), 4) Physical Functioning scale, 5) Role Limitations-Physical scale of the SF-36, and 6) The Main Complaint (MC). Subjects ($n = 97$) that still suffered residual complaints 6 weeks following a lumbar disc surgery completed the questionnaires before and 3 months after treatment. In a direct comparison the A) The test-retest reproducibility (Intraclass Correlation Coefficients [ICC] and the Standard Error of Measurement [SEM]) and B) 3 parameters of responsiveness (Minimal Detectable Change [MDC], Standardised Response Mean [SRM], and the Area Under the receiver operator characteristic Curve [AUC]) were assessed.

Results: This study suggests the superiority of the 3 versions of the RDQ compared to the 3 other questionnaires. Comparing the 3 versions of the RDQ reveals no substantial differences thereby indicating that the 2 modified version of the RDQ hold no better measurement properties in this specific population.

Conclusion: The use of the RDQ-24 for this specific post-surgery population is suggested. The optimal cut-off point of the RDQ-24 that minimizes the overall classification error was found to be 3.5 with a sensitivity of 94.6% and a specificity of 88.2%.

Keywords: Functional status; Low back pain; Lumbar disc surgery; Roland-Disability Questionnaire; SF-36; Measurement properties

1. Introduction

A recent systematic review that assessed the effectiveness of various rehabilitation programs for patients after lumbar disc surgery concluded that is was unclear what the exact content of active post-surgery rehabilitation should be [1]. The aim of rehabilitation programs in general is to improve the functional status of patients. Restriction of function is a patient-referenced concept that is different for each individual. Questionnaires have been developed for measuring functional status, but they were designed for patients with nonspecific low back pain. One widely used questionnaire is the Roland-Morris Disability questionnaire (RDQ) [2]. Several modifications to the RDQ have been suggested, but these modifications seem to provide only modest improvement in measurement properties that were in general considered to be satisfactory [3]. An international group of experts [4] suggested the use of the original version because in addition to the satisfactory measurement properties the RDQ has been widely used in many studies (and countries) for nonspecific low back pain.

However, measurement properties of questionnaires should be based on samples of populations on whom these measurements will be used in clinical practice. Davidson and Keating [5] argue that differences in measurement properties of two versions of the RDQ between their study and the study by Patrick et al. [6] can be due to differences in study populations. Whereas Davidson and Keating included patients that were seeking help from a physical therapist, Patrick et al. included patients with sciatica secondary to a herniated lumbar intervertebral disk. The measurement properties of the various versions of the RDQ for patients after a lumbar disc surgery have not been assessed. In addition to the various versions of the RDQ, the measurement

* Corresponding author. Tel.: +31-20-4448149; fax: +31-20-4448181.
E-mail address: r.ostelo@vumc.nl (R.W.J.G. Ostelo).

R.W.J.G. Ostelo worked at the Department of Epidemiology, Maastricht University while conducting the randomized controlled trial that was the basis for this study.

0895-4356/04/$ – see front matter
doi: 10.1016/j.jclinepi.2003.09.005

properties of other, widely used questionnaires, such as the Physical Functioning subscale and the Role Limitations-Physical subscale of the SF-36 [7] and the Main Complaint (MC) [8], are unknown in this specific population. The SF-36 subscales and the MC were chosen because these are relatively brief instruments. Several authors have advocated the direct comparisons of evaluative functional status questionnaires in a single patient group [5,9,10].

The purpose of this study is to establish an evidence-based recommendation regarding the use of functional status questionnaires in patients after lumbar disc surgery by a direct comparison of the reproducibility and responsiveness of the included functional status questionnaires. Reproducibility is defined as the ability to measure attributes in a consistent manner when administered on several occasions to stable subjects [11]. In other words, the question to be answered for evaluative instruments is whether the instrument produces similar results on repeated administration when no real change in health status has occurred within this time frame [12]. Lack of reproducibility can be due to random measurement error [11] and real within-subject variance [13,14]. Both components together lead to measurement fluctuations in health status in the absence of real change. These fluctuations have been termed "background noise" [12].

Responsiveness is the extent to which different results are obtained on repeated administration of the same instrument when a real change in health status has occurred [13,15–17]. There are various ways of expressing the responsiveness, and there is no consensus on the most appropriate strategy for quantifying responsiveness. In this article, responsiveness is defined as the ability to detect clinically important changes [11,18]. To explore the responsiveness from this perspective, an explicit external criterion is used to define whether a patient has deteriorated, has not changed, or has improved such that the improvement can be considered clinically relevant.

2. Methods

2.1. Study population

The study population consisted of participants ($n = 105$) of a randomized controlled trial concerning the effectiveness on the rehabilitation after lumbar disc surgery [19,20]. Patients were checked for eligibility for this trial by the neurosurgeon during the routine 6-week post-surgery consultation. Patients were eligible if they were between 18 and 65 years of age, if this was their first-time lumbar disc surgery, and if they still were restricted in the activities of daily life or had not yet (fully) resumed work 6 weeks after surgery. Patients with a confirmed and relevant co-morbidity that possibly affected the spine, such as morbus Bechterew, were excluded. Patient characteristics were collected at baseline. Follow-up measurement took place immediately after the treatment period (3 months after randomization). The medical ethics committee of the Maastricht University Hospital approved the research protocol.

2.2. Questionnaires

2.2.1. Functional status questionnaires

The following functional status questionnaires, completed by the patients, were taken into account for the purpose of this study. The same research assistant was present during both measurements (at baseline and after 3 months) in case patients had difficulties completing the questionnaires. Because these analyses are based on data obtained from the trial, the choice of the competing measures was determined by the design of the trial. In the trial, the original Roland-Morris Disability questionnaire (RDQ-24) was selected as main outcome measure because Beurskens [21] reported higher point estimates of change and an international group of experts [4] suggested the use of the original RDQ. Two modifications of the RDQ were administered to analyze whether these modifications had better measurement properties. We also assessed whether other questionnaires measuring functional status, as listed below, had better measurement properties as compared with the RDQ-24.

1. The original (RDQ-24) [2] contains 24 yes/no items. Patients are asked whether the statements apply to them that day (the last 24 hours). The RDQ-24 score is calculated by adding up the number of "yes" items, ranging from 0 (no disability) to 24 (maximum disability). Because in this population we expected sciatica to be prevalent, we changed the terminal phrase of each statement from "because of my back pain" to "because of my back or leg problem" as suggested by Patrick [6].
2. The Modified Roland-Morris Disability questionnaire (MRDQ) contains 23 items. Patrick [6] suggested that the responsiveness of the RDQ-24 could be increased by removing five potentially redundant items and adding four additional items relating to sexual function, daily work, expressions of concern to others, and the need to rub or hold areas that hurt.
3. The short Roland-Morris Disability questionnaire (RM-18) contains 18 items because Stratford [22] concluded that six items were redundant and measurement properties were equivalent to the RDQ-24.
   (Both modifications [MRDQ and RM-18] used the phrasing "because of my back or leg problem." Scoring methods were also identical to the RDQ-24.)
4. The 10-item Physical Functioning scale of the SF-36 (version 1) [7] (SF-36 PhF) is used to measure activity limitations experienced at *this* moment. Every item has three response options. After transformation of the scores, the range is from 0 (maximum limitations) to 100 (no limitations).
5. The four-item Role Limitations-Physical scale of the SF-36 (version 1) [7] measures activity limitations experienced the last 4 weeks. Every item has two

response options. After transformation of the scores, the range is from 0 (maximum limitations) to 100 (no limitations).

6. The Main Complaint (MC) [8] measures the limitation that patients experience while performing activities selected at baseline in a standardized way. Patients selected three activities they performed frequently, that they perceived as important in their daily life, and that were hampered by their back or leg complaints. Patients rated the severity of these main complaints on a 100-mm visual analogue scale. For this study, only the first main complaint was used.

2.3. External criterion

In this study, the explicit external criterion exploring the responsiveness was the seven-point global perceived effect (GPE) scale (1 = completely recovered, 7 = worse than ever) that was administered at the 3-month follow-up. If a patient indicated "complete recovery" or "much improved," the patient was coded as "improved," which was regarded as clinically important. Patients who indicated "slightly improved," "no change," or "slightly worsened" were coded as "unchanged." Patients who rated their complaints as "much worsened" or "worse than ever" were coded as "deteriorated." This is concordance with Beurskens [10].

2.4. Data analysis

2.4.1. Reproducibility

Patients who reported to be unchanged were included in the test-retest analysis, assuming that these patients had no clinically relevant improvement. Because the large number of statements that patients had to respond to and the 3-month interval, it was assumed that patients were not able to recall their initial responses.

Test-retest reproducibility was assessed using the following two methods.

1. In a two-way random model, the intraclass correlation coefficients (ICC) for agreement [23] were calculated for each questionnaire as the ratio of variance between subjects and the total variance. These variance components were computed with ANOVA for random effects. The ICC ranges from 0 to 1.
2. The standard error of measurement (SEM) was calculated to express measurement error in the same units as the original questionnaire. The SEM is defined as the square root of the within-subject variance consisting of variance between measures (to account for systematic error between measurements) and the residual variance. The 95% confidence interval (CI) was calculated as described by Brennan [24].

2.4.2. Responsiveness

The responsiveness of each questionnaire was investigated in the following three ways:

1. The SEM (based on "unchanged" subjects) was used to calculate the minimal detectable change (MDC). The MDC is calculated as $1.96 \times \sqrt{2} \times \text{SEM}$. Many authors state that the MDC expresses the minimal magnitude of change, expressed in scale points, required to be 95% confident (hence the 1.96) that the observed change between the two measures (hence the $\sqrt{2}$) reflects real change and not just measurement error. Because only stable patients are assessed, one does not know the likelihood that a patient has truly changed. Therefore, a more accurate interpretation of the MDC is that it expresses the magnitude of change, with a chance of less than 5%, that a patient being stable is truly stable. Given this small probability, it is likely that a patient whose score exceeds the MDC has changed. Because the explicit external criterion that is used labels patients as deteriorated, not changed or as improved such that the improvement can be considered clinically relevant, in this study the MDC can be interpreted as clinically important change.
2. Another method to investigate responsiveness is relating the magnitude of change to the variability in score [25]. The effect size statistic is calculated by taking the mean change found in a variable (the signal) and dividing it by the standard deviation (SD) of that variable (the noise). In the clinimetric literature, there is no consensus with regard to what SD to take. Some authors propose the SD of the baseline scores [18,26]; the statistic is then referred to as effect size. Others have suggested taking the SD of the mean change score of the same group [5,10,25,27]. This statistic often is referred to as standardized response mean (SRM). In this study, we calculated the SRM because we believe that measuring change is a function of the SD of the change score. SRMs were calculated for "unchanged" and "improved" subjects separately.
3. A third method of evaluating responsiveness assesses the ability of an instrument to discriminate between clinically relevant (improved) and clinically irrelevant changes (unchanged). If the functional questionnaires are considered as diagnostic tests for improvement, then these instruments can be described in terms of sensitivity and specificity for detecting change as established by the gold standard [13]. The receiver operator characteristic (ROC) curve is a graph of true positive (sensitivity) versus false positive (1-specificity) for each of several cut-off points in score change. The area under the ROC curve reflects the ability of the test to discriminate between subjects who have improved from subjects who are unchanged. A value of 1 for the AUC represents perfect (100%) accuracy, whereas a value of 0.5 represents chance alone [13]. For every questionnaire, a cut-off point is calculated for which sensitivity and specificity jointly minimize the total error in misclassification.

A prerequisite for measuring improvement is a prior score that does not equal the best possible score on a particular scale; the same holds true for measuring deterioration. In other words, prior scores at the floor or at the ceiling of the scale affect the ability of the scale to detect a meaningful change. Davidson and Keating [5] have coined the term "scale width" to indicate the capacity of a scale to have initial scores that are far enough onto the scale to allow detection of change in scores over time. Because we defined the MDC as the minimal required change, we examined scale width in terms of not more than 15% of the respondents within 1 MDC from the theoretical minimum or maximum of a particular scale. Scale width as defined here assumes a single measurement. Averaging measurements over occasions can reduce scale width. For all statistics, SPSS 10.1 for Windows was used.

To summarize the results, we tried to define criteria regarding the various measurement properties investigated. The scale width was only measurement property for which we found criteria in the literature. For scale width, the 15% criterion was suggested by McHorney et al. [28]. If less than 15% of the responders had initial scores within 1 MDC from one anchor of the scale, this was labeled "good"; scale width was labeled "negative" if initial scores exceeded the 15% rule. The ability of measuring improvement and deterioration were assessed separately. Our interpretation of the standardized response means (SRMs) was broadly based on the criteria for the effect size as described by Cohen [25]: $\leqslant 0.20$ is regarded as "negative," >0.20 and $\leqslant 0.50$ as "doubtful," >0.5 and <0.8 as "good," and values of 0.80 or greater as "very good." These values apply to the SRM for "improved" subjects, whereas for "unchanged" subjects SRM values had to be <0.50 to be rated "good," and values of $\geqslant 0.50$ were rated as "negative." For the ICC, we consider values $\leqslant 0.60$ as "negative," >0.60 and $\leqslant 0.80$ as "doubtful," >0.80 and <0.90 as "good," and 0.90 or greater as "very good." For the AUC, the same criteria were used. For the SEM, we calculated the corresponding percentage of SEM related to the total score of the scale. Criteria that we used were: $\leqslant 5\%$ was "very good," $>5\%$ and $\leqslant 10\%$ was "good," $>10\%$ and $<20\%$ was "doubtful," and values of 20% were considered as "negative". Because the MDC is based on the SEM, no criteria for the MDC were defined.

3. Results

Of the 105 patients included in the trial, eight dropped out before the 3 months post-treatment measurement. Table 1 presents the information on the characteristics of the 97 participants with complete data on pre-treatment and post-treatment measurements.

The mean scores at pre-treatment and the mean scores at post-treatment for subjects in each of the seven categories of GPE, for the total group, and for the two clustered groups that were used in the analyses are presented in Table 2.

Table 1
Characteristics of patients ($n = 97$)

Variable	
Mean age in years (SD)	43.3 (8.8)
Sex (% female)	41.2%
Duration of this episode before operation	
2–6 months	51.5%
7–12 months	39.2%
13 months or more	9.3%
Previous low back or leg pain (%)	81.4%
Work status	
Employed	80.4%
Unemployed	5.2%
Not in labor force	14.4%

On the seven-point GPE, five patients rated themselves as "completely recovered," and 51 patients rated themselves as "much improved." Only seven patients indicated deterioration ("much worsened" or "worse than ever"). Due to their small number, they were excluded from the analysis. These results show that the direction and the magnitude in differences between baseline measurement and post-treatment measurement are as expected: The change over time for each questionnaire declined with the decreasing categories of GPE. Moreover, the results showed that there were no statistically significant differences between the change scores of the "slightly improved" group and the "no change" group. However, there were statistically significant differences between the change scores of the "slightly improved" group and the "much improved" group.

Table 3 presents all ICCs. The ICCs ranged from 0.14 (95% confidence interval [CI] 0–0.40) for the MC to 0.78 (95% CI 0.57–0.89) for the MRDQ. The SEM, the MDC, and the between-subject and within-subject variance components for all questionnaires are also presented in Table 3.

Table 4 shows that for the subjects that were "unchanged" ($n = 34$), the mean change scores were small, indicating no difference between pre-treatment and post-treatment measurement except on the MC and on the SF-36 RLPh, where scores improved by 22.2 mm (SD 24.1 mm) and 14.7 (SD 32.6), respectively, between initial measurement and post-treatment.

Table 4 shows that the mean changes of the questionnaires differed between the patients in the "improved" and "unchanged" group. SRMs in the "improved" group exceed the SRMs in the "unchanged" group. The SRMs in the improved group could all be labeled as "very good." The RDQ24 and the MC showed the largest SRMs (2.02), whereas the SRM of the SF-36-RLPh was the smallest with 1.40. The areas under the ROC curve (AUC) showed that there were no differences between the three versions of the RDQ as indicated by the largely overlapping 95% CIs (Fig. 1). The three remaining questionnaires had smaller AUC, especially the SF-36 RLPh and the MC.

Assessing the scale width (Table 5) revealed that the three versions of the RDQ were comparable with regard to detecting improvement. For detecting deterioration, MRDQ

Table 2
Mean scores and standard deviations[a] at pre-treatment (T0) and at post-treatment (T1) for subjects per category of GPE for the total group and for collapsed categories as used in analysis: "improved" and "unchanged"

	RDQ-24		MRDQ		RDQ-18		SF36-PhF		SF36-RLPh		MC	
	T0	T1	T0	T1	T0	T1	T0	T1	T0	T1	T0	T1
Categories of GPE												
Completely recovered ($n = 5$)	15.4 (2.1)	1.0 (2.2)	14.6 (2.1)	0.8 (1.8)	14.0 (2.3)	0.8 (1.8)	59.0 (14.7)	92.0 (7.6)	5.0 (11.2)	70.0 (44.7)	66.2 (20.3)	13.8 (12.8)
Much improved ($n = 51$)	12.8 (4.3)	4.6 (4.7)	12.7 (4.3)	4.6 (5.0)	10.9 (3.6)	4.0 (4.2)	56.5 (16.4)	81.4 (12.7)	5.9 (12.8)	61.3 (41.9)	65.6 (16.1)	24.8 (18.6)
Slightly improved ($n = 23$)	15.3 (3.6)	13.9 (4.2)	15.7 (3.8)	14.2 (4.2)	13.4 (3.3)	12.3 (3.7)	48.9 (18.5)	57.6 (14.1)	7.6 (17.6)	21.7 (34.8)	74.4 (12.6)	49.5 (22.0)
No change ($n = 9$)	15.7 (4.5)	15.0 (3.5)	15.7 (4.0)	15.2 (3.0)	13.1 (3.1)	12.9 (2.5)	45.0 (12.7)	45.7 (21.0)	0.0 (0.0)	16.7 (33.1)	78.4 (13.9)	66.6 (28.2)
Slightly worsened ($n = 2$)	15.5 (2.1)	15.0 (1.4)	16.0 (4.2)	16.5 (3.5)	13.5 (2.1)	13.0 (1.4)	25.0 (0.0)	50.0 (7.1)	0.0 (0.0)	12.5 (17.7)	90.0 (14.1)	53.5 (3.5)
Much worsened ($n = 6$)	14.3 (4.0)	15.8 (3.7)	14.5 (4.1)	16.0 (4.6)	12.2 (3.7)	13.5 (3.4)	45.8 (16.9)	35.8 (21.3)	16.7 (20.4)	0.0 (0.0)	63.0 (16.1)	69.9 (32.7)
Worse than ever ($n = 1$)	22.0	23.0	22.0	23.0	18.0	18.0	5.0	0.0	0.0	0.0	75.0	82.0
Total group ($n = 97$)	14.0 (4.2)	8.7 (6.8)	14.1 (4.3)	8.8 (7.0)	12.1 (3.6)	7.6 (5.9)	51.9 (17.8)	68.7 (22.6)	6.2 (14.0)	36.6 (43.5)	74.5 (19.2)	43.4 (29.8)
Categories used in analysis												
Improved ($n = 56$)	13.0 (4.2)	4.3 (4.7)	12.9 (4.2)	4.3 (4.9)	11.2 (3.6)	3.8 (4.2)	56.7 (16.1)	82.3 (12.7)	5.8 (12.6)	62.1 (41.8)	70.8 (19.7)	26.6 (19.7)
Unchanged ($n = 34$)	15.4 (3.7)	14.3 (3.9)	15.7 (3.7)	14.6 (3.8)	13.4 (3.1)	12.5 (3.3)	46.5 (17.3)	54.0 (16.4)	5.1 (14.8)	19.9 (33.0)	80.9 (16.8)	65.0 (25.5)

Abbreviations: GPE, global perceived effect; RDQ-24, Roland-Morris Disability Questionnaire; MRDQ, Modified Roland-Morris Disability Questionnaire; RDQ-18, Roland-Morris Disability Questionnaire 18 items; SF-36 PhF, Physical Functioning scale SF-36; SF-36 RLPh, Role Limitations-Physical scale SF-36; MC, Main Complaint.

[a] Standard deviations shown in brackets.

slightly exceeded the 15% rule, whereas RDQ-18 performed best with 0%. The SF-36 RLPh and the MC exceeded the 15% rule by far with regard to detecting deterioration. Table 6 presents a summary of the results

4. Discussion

In this study, the reproducibility and responsiveness of the six questionnaires measuring functional status in patients after a lumbar disc surgery were assessed in a direct comparison. The reproducibility was assessed in the test-retest analysis including "unchanged" patients. The ICC values of the three versions of the RDQ were comparable, ranging from 0.74 (RDQ-24) to 0.78 (MRDQ). The magnitude of these ICCs was labeled as "doubtful." The ICCs of the other three questionnaires were all labeled as "negative." The values of the SEM showed that the three versions of the RDQ were similar and superior to the other three functional status questionnaires. This indicates that none of the questionnaires included in this project was superior to the RDQ-24.

Responsiveness was defined as the ability to detect clinically important changes. For all three versions of the RDQ, the MDCs were smaller than the other three questionnaires, meaning that these are more sensitive in detecting a clinically important change. The MDC (expressed in percentages of the scale range) for the MC and the SF-36 RLPh were 63.6% and 69.3%, respectively. The SRM was calculated and the AUC were assessed to evaluate responsiveness in terms of sensitivity to change and specificity to change. The results of both strategies led to the same conclusion: The three versions of the RDQ performed better in discriminating between "improved" and "unchanged" as compared with both the subscales of the SF-36 or the MC. The SRM of the main complaint for improvement was comparable to the SRM of the RDQ-24; however, also in "unchanged" patients, the SRM of the MC was large. Most likely the MC is measuring something other than or in addition to the outcome of interest. Finally, because of floor and ceiling effects, the MC and the SF-36 RLPh might have serious difficulties in detecting deterioration, whereas the MRDQ only slightly exceeds the 15% rule for detecting deterioration.

Results from this study suggest that the three versions of the RDQ are superior to the two subscales of the SF-36 and the MC regarding the reproducibility and the responsiveness. Comparing the three versions of the RDQ reveals only small differences. The MRDQ has a slightly higher ICC value. The SEMs, and consequently the MDCs, are comparable, but the MRDQ slightly exceeds the 15% rule at the upper end of the scale. The RDQ-24 is slightly better in discriminating between "unchanged" and "improved" as indicated by the values of the SRMs and the AUCs.

The measurement properties were assessed in this specific post-surgery population because measurement properties are not fixed numbers but are highly dependent on the patient group, including diagnosis and stage and the timing of data

Table 3
The variance components and indexes

Questionnaire	Between subject variance	Within-subject variance		ICC (95% CI)	SEM (95% CI)	SEM (%)[a] (95% CI)	MDC[b] (95% CI)	MDC (%)[c] (95% CI)
		Between measures	Residual					
RDQ-24	11.152	0.596	3.257	0.74 (0.51–0.87)	2.0 (1.5–2.9)	8.2 (6.3–12.1)	5.4 (4.2–8.0)	22.5
MRDQ	11.520	0.512	2.708	0.78 (0.57–0.89)	1.8 (1.4–2.6)	7.2 (5.6–10.4)	5.0 (3.9–7.2)	21.7
RDQ-18	7.868	0.271	2.317	0.75 (0.55–0.87)	1.6 (1.2–2.0)	8.9 (6.7–11.1)	4.5 (3.3–5.5)	25.0
SF-36 PhF	185.660	25.561	98.442	0.60 (0.28–0.79)	11.1 (8.2–17.4)	11.1 (8.2–17.4)	30.9 (22.7–48.2)	30.9
SF-36 RLPh	121.992	92.469	532.531	0.16 (0–0.45)	25.0 (22.8–27.4)	25.0 (22.8–27.4)	69.3 (63.2–75.9)	69.3
MC	83.597	237.373	289.832	0.14 (0–0.40)	23.0 (14.0–61.2)	23.0 (14.0–61.2)	63.6 (38.8–100)	63.6

Abbreviations: ICC, intraclass correlation coefficient for agreement based on a two-way random effect model; RDQ-24, Roland-Morris Disability Questionnaire; MRDQ, Modified Roland-Morris Disability Questionnaire; RDQ-18, Roland-Morris Disability Questionnaire 18 items; SF-36 PhF, physical functioning scale SF-36; SF-36 RLPh, role limitations-physical scale SF-36; MC, Main Complaint; SEM, standard error of measurement ($\sqrt{}$ within-subject variance).

[a] SEM (%) is SEM expressed in percentages of corresponding scale.

[b] MDC (minimal detectable change in scale points) = $1.96 \times \sqrt{2} \times \text{SEM}$.

[c] MDC (%) is MDC expressed in percentages of corresponding scale.

collection. However, our study shows that the ICC and the MDC of the original RDQ (RDQ-24) in this population are comparable with studies including patients with chronic low back pain [3]. There have been several proposals for modifications of the original RDQ. These modifications seem to provide only modest improvement in measurement properties that were in general considered to be satisfactory [3], and an international group of experts [4] suggested the use of the original RDQ version because, in addition to the satisfactory measurement properties, the original RDQ has been widely used in many countries for nonspecific low back pain. Results from our study show that both modifications, as suggested by Patrick et al. [6] and by Stratford et al. [22], did not lead to significantly better results with regard to the reproducibility or responsiveness as compared with the RDQ-24. This is perhaps not surprising because there is a large overlap in items. Therefore, we suggest also the use of the RDQ-24 for this specific post-surgery population because of its satisfactory measurement properties and because of its ability to enhance the comparison between studies.

We did not test whether the various measures for reproducibility and responsiveness statistically significantly differed for the included questionnaires. The outcomes for the three versions of the RDQ seem similar, which indicates that the two modified versions of the RDQ did not perform better. Moreover, it is important to include all psychometrical measurement, instead of looking at one specific measure. Therefore, we summarized the results in a descriptive way (see Table 6). Regarding this way of summarizing, it is questionable whether the cut-off points we used for the various defined categories (e.g., "very good" or "good") to summarize the results of this study were correctly chosen. The cut-off points for the defined are, even if based on literature, arbitrary. The advantage of this procedure is that defining criteria and cut-off points explicitly makes the procedure of how the conclusions were reached transparent. Moreover, by presenting the criteria explicitly, readers are able to draw their own conclusions. Modifying the cut-off points hardly changed the results of this study: the three versions of the RDQ were superior to the other questionnaires. Moreover, the results for the three versions of the RDQ are similar, which indicates that the modifications do not perform better in this specific population.

Table 4
Mean change scores and standard deviations (SD) and standardized response mean (SRM) for subjects classified as "unchanged" and "improved" (all scores are calculated such that positive scores mean improvement for patient)

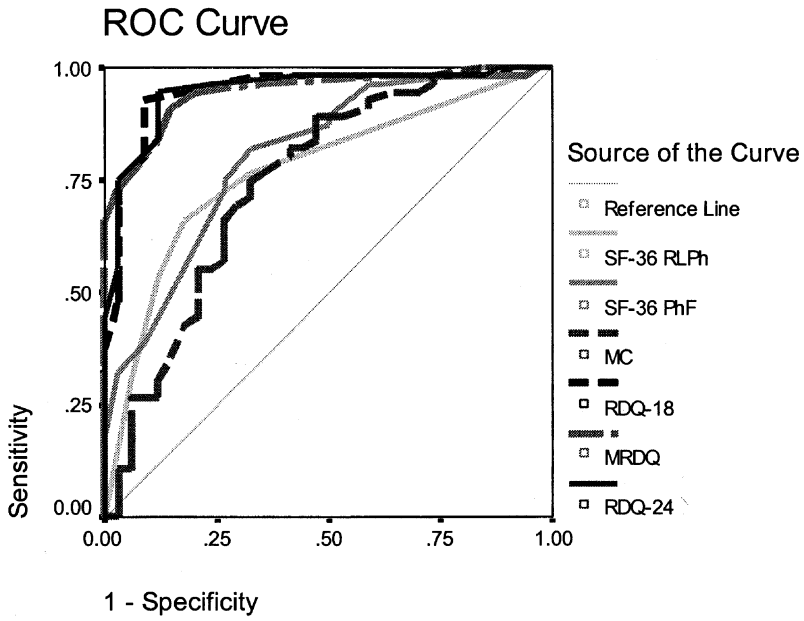
Changed	RDQ-24	MRDQ	RDQ-18	SF-36 PhF	SF-36 RLPh	MC
Unchanged ($n = 34$)						
Mean change	1.2	1.1	0.8	7.5	14.7	22.2
SD	2.6	2.3	2.2	14.0	32.6	24.1
SRM[a]	0.46	0.48	0.36	0.54	0.45	0.92
Improved ($n = 56$)						
Mean	8.7	8.6	7.4	25.6	56.3	41.9
SD	4.3	4.5	3.9	14.4	40.0	20.7
SRM	2.02	1.91	1.90	1.78	1.40	2.02

Abbreviations: RDQ-24, Roland-Morris Disability Questionnaire; MRDQ, Modified Roland-Morris Disability Questionnaire; RDQ-18, Roland-Morris Disability Questionnaire, 18 items; SF-36 PhF, Physical Functioning Scale SF-36; SF-36 RLPh, Role Limitations-Physical Scale SF-36; MC, Main Complaint; SRM, standard response mean.

[a] SRM is calculated by dividing the mean change score by the SD of the mean change score.

4.1. External criteria for measuring change

In this study, we were interested in change in functional status. We used global perceived effect as the external criteria. Our clustering into "improved" and "unchanged" groups was in concordance with Beurskens [10]. There was support for including "slightly improved" in the "unchanged" group because there were statistically significant differences between the change scores of the "slightly improved" group and the "much improved" group. Moreover, there were no statistically significant differences between the change scores



Area under the Curve

Questionnaires (Change scores)	Area	Std. Error	95% CI [a]		Cut-off		
			LB	UB		Se	Sp
RDQ-24	.949	.024	.902	.995	3.5	94.6	88.2
MRDQ	.944	.024	.898	.991	3.5	91.1	85.3
RDQ-18	.948	.025	.899	.997	2.5	92.9	91.2
SF-36 PhF	.810	.047	.719	.902	12.5	82.1	67.6
SF-36 RLPh	.773	.051	.673	.873	37.5	66.1	82.4
MC	.752	.056	.643	.861	29.5	75.0	67.6

RDQ-24 = Roland-Morris Disability Questionnaire [2], MRDQ = Modified Roland-Morris Disability Questionnaire [6], RDQ-18 = Roland-Morris Disability Questionnaire 18 items [22], SF-36 PhF = Physical Functioning scale SF-36, SF-36 RLPh = Role Limitations-Physical scale SF-36, MC = Main Complaint[8]. 95% CI = 95% confidence interval with LB (lower bound) and UP (upper bound). Cut-off points are defined as change scores in points on the original scales in terms of improvement for patients. The criterion was the minimal total error in misclassification. Se (sensitivity), Sp (specificity). [a] 0-hypothesis is Area = 0.5

Fig. 1. ROC curves and areas under the curve.

Table 5
Scale width of questionnaires at baseline

Questionnaire	Percentage of subjects with initial scores that hamper detection of improvement	Percentage of subjects with initial scores that hamper detection of deterioration
RDQ-24	0	10
MRDQ	0	17
RDQ-18	0	0
SF-36 PhF	12	10
SF-36 RLPh	1	99
MC	8	83

Abbreviations: RDQ-24, Roland-Morris Disability Questionnaire; MRDQ, Modified Roland-Morris Disability Questionnaire; RDQ-18, Roland-Morris Disability Questionnaire, 18 items; SF-36 PhF, Physical Functioning scale SF-36; SF-36 RLPh, Role Limitations-Physical Scale SF-36; MC, Main Complaint.

of the "slightly improved" group and the "no change" group, both of which were included in the "unchanged" group. Still, regarding the use of this external criterion, some comments have to be made. First, there is the issue of correlated error because every patient completes both the included questionnaires and the external criteria. Furthermore, Norman et al. [29] question the validity of these single-item global rating of unknown reproducibility as the standard for evaluating a multi-item tool that presumably yields measurements of superior reproducibility compared with the global rating. Their concern is that global ratings may be influenced by recall bias, for which there is some evidence [30]. Moreover, global perceived effect, as used in the current study, is an all-encompassing measure for improvement that includes

Table 6
Summary of the results of six evaluative functional status questionnaires

	Reproducibility			Responsiveness	
Questionnaire	ICC	SEM	SRM (improved/unchanged)	AUC	Scale width (improvement/deterioration)
RDQ-24	+/−	+	++/+	++	+/+
MRDQ	+/−	+	++/+	++	+/−
RDQ-18	+/−	+	++/+	++	+/+
SF-36 PhF	−	+/−	++/−	+/−	+/+
SF-36 RLPh	−	−	++/+	+/−	+/−
MC	−	−	++/−	+/−	+/−

Abbreviations: ICC, intraclass correlation coefficient; SEM, standard error of measurement; SRM, standard response mean; AUC, area under the curve; RDQ-24, Roland-Morris Disability Questionnaire; MRDQ, Modified Roland-Morris Disability Questionnaire; RDQ-18, Roland-Morris Disability Questionnaire, 18 items; SF-36 PhF, Physical Functioning Scale SF-36; SF-36 RLPh, Role Limitations-Physical Scale SF-36; MC, Main Complaint.
++, very good; +, good; +/−, doubtful; −, negative.

pain, functional status, and other aspects that patients perceive as important. This is no gold standard that defines whether a patient's functional status has changed. However, from the patients' and the clinicians' viewpoints, it is relevant and sensible to ask the patient to assess his or her perceived benefit [31,32]. We consider it a surrogate criterion. With regard to the responsiveness, this implies that this criterion does not precisely define the smallest amount of change that is clinically relevant. Consequently, the background noise estimation based on this surrogate criterion and the real background noise typically differ. Therefore, it is not possible to assess absolute responsiveness. However, this surrogate external criterion can be used to compare the various measures of reproducibility and responsiveness in a direct comparison as presented in this study [12,27,33–36]. Some authors have suggested that more comparisons of functional status measures against several external criteria should be analyzed because if results are consistent on the basis of several criteria, confidence increases about the correct ranking of the measures [10,37]. However, researchers should be aware that various external criteria might reflect different perspectives: the patient's, clinician's, payer's, or society's perspective reflect different concepts of functional status. If these various perspectives or concepts are used for estimating when a patient is "better," this may well correspond to different amount of change [38].

4.2. Various methods for assessing reproducibility

For assessing the reproducibility, we used two methods: the ICC and SEM. The ICC is generally accepted in the medical literature as the preferred method for quantifying reliability [11,18,37,39]. We calculated the ICC for agreement (as opposed to the ICC for consistency) [23] because we view differences between the two measurements in absolute scores on a questionnaire, regardless of the reason, as disagreement. However, the variance of interest with regard to the ICC is the between-subjects measures, whereas in longitudinal changes the magnitude of the within-subject variance over time is relevant [15,40,41]. Furthermore, the ICC, expressed as a dimensionless number between 0 and 1, is hardly interpretable in terms of scale points. In addition, Beckerman et al. [40] have demonstrated that reliability coefficients such as the ICC are not appropriate for gaining insight into the methodologic quality of instruments measuring change over time within a subject. Because the SEM and the related MDC are better suited for that purpose [40], these measures are presented in this study. However, the SEM and the related MDC have many faces, and even the taxonomy varies considerably [42]. Because the design of this study follows a two-way random effects model, the SEM was calculated as the square root of the within-subject variance consisting of the variance between measures and the residual variance (see Table 3). In this way, systematic disagreements between the two measurements are accounted for.

4.3. Various methods for evaluating responsiveness

Three strategies have been used to assess responsiveness. We considered the MDC, based on "unchanged" subjects as a measure for responsiveness, because it defines the minimal change needed for labeling the change (with 95% confidence) as a clinically relevant change. In the present study, the MRDQ (21.7%) and the RDQ-24 (22.5%) seemed to be most sensitive to detect clinically important change. The two other strategies (SRM and ROC curves) have their advantages. The advantage of the SRM is that it is easier to calculate, although there is the controversy with regard to what SD to use. The ROC curves visualize the relation between the true-positive and false-positive rates at different cut-off points of change scores. Moreover, the ROC curve identifies the optimal cut-off point for the desired combination of sensitivity and specificity.

In conclusion, the measurement properties of the three versions of the RDQ were similar and superior to the other three functional status questionnaires, thereby indicating that none of the other included questionnaires holds better measurement properties than the RDQ-24. Based on the results of this study, we suggest the use of the RDQ-24 for this specific post-surgery population. The optimal cut-off point of the RDQ-24 that minimizes the overall classification error was found to be 3.5, with a sensitivity of 94.6% and a specificity of 88.2%.

References

[1] Ostelo RW, de Vet HC, Waddell G, Kerckhoffs MR, Leffers P, van Tulder MW. Rehabilitation after lumbar disc surgery: a systematic review within the framework of the Cochrane Collaboration. Spine 2003;28:209–18.
[2] Roland M, Morris R. A study of the natural history of low-back pain, Part 2: development of guidelines for trials of treatment in primary care. Spine 1983;8:145–50.

[3] Roland M, Fairbank J. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine 2000;25:3115–24.
[4] Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. Spine 2000;25:3100–3.
[5] Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 2002;82: 8–24.
[6] Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995;20:1899–908.
[7] Ware Jr. JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I: conceptual framework and item selection. Med Care 1992;30:473–83.
[8] Beurskens AJHM, de Vet HCW, Köke AJ, Lindeman E, van der Heyden GJ, Regtop W, Knipschild PG. A patient-specific approach for measuring functional status in low back pain. J Manipulative Physiol Ther 1999;22:144–8.
[9] Garratt AM, Klaber Moffett J, Farrin AJ. Responsiveness of generic and specific measures of health outcome in low back pain. Spine 2001;26:71–7.
[10] Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71–6.
[11] Streiner DL, Norman GR. Health measurement scales. 2nd edition. Oxford, New York, Tokyo: Oxford University Press; 1995.
[12] de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures: theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care 2001;17:479–87.
[13] Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39:897–906.
[14] Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38:27–36.
[15] Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992;45: 1341–5.
[16] Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992;1:73–5.
[17] Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369–78.
[18] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40: 171–8.
[19] Ostelo RWJG, Köke AJA, Beurskens AJHM, Kerckhoffs MR, Vlaeyen JWS, Wolters PM, Berfelo MW, van den Brandt PA. Behavioral-graded activity compared with usual care after first-time disk surgery: considerations of the design of a randomized clinical trial. J Manipulative Physiol Ther 2000;23:312–9.
[20] Ostelo RWJG, de Vet HCW, Vlaeyen JW, Kerckhoffs MR, Berfelo MW, Wolters PM, van den Brandt PA. Behavioral graded activity after first-time lumbar disc surgery: one year results of a randomized controlled trial. Spine 2003;28:1757–65.
[21] Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain: assessment of the quality of four disease-specific questionnaires. Spine 1995;20:1017–28.
[22] Stratford PW, Binkley JM. Measurement properties of the RM-18: a modified version of the Roland-Morris Disability Scale. Spine 1997; 22:2416–21.
[23] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
[24] Brennan JL. Generalizability theory. New York: Springer; 2001.
[25] Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1977.
[26] Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178–89.
[27] Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79–93.
[28] McHorney CA, Ware JE Jr. Construction and validation of an alternate form general mental health scale for the Medical Outcomes Study Short-Form 36-Item Health Survey. Med Care 1995;33:15–28.
[29] Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869–79.
[30] Linton SJ, Melin L. The accuracy of remembering chronic pain. Pain 1982;13:281–5.
[31] Bombardier C, Tugwell P, Sinclair A, Dok C, Anderson G, Buchanan WW. Preference for endpoint measures in clinical trials: results of structured workshops. J Rheumatol 1982;9:798–801.
[32] Fries JF. Toward an understanding of patient outcome measurement. Arthritis Rheum 1983;26:697–704.
[33] van der Heijden GJ, Leffers P, Bouter LM. Shoulder disability questionnaire design and responsiveness of a functional status measure. J Clin Epidemiol 2000;53:29–38.
[34] Bronfort G, Bouter LM. Responsiveness of general health status in chronic low back pain: a comparison of the COOP charts and the SF-36. Pain 1999;83:201–9.
[35] van der Windt DA, van der Heijden GJ, de Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. Ann Rheum Dis 1998;57:82–7.
[36] Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther 1998;78:1186–96.
[37] Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Control Clin Trials 1991;12:142S–58S.
[38] Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine 2000;25:3192–9.
[39] de Vet HCW. Observer reliability and agreement. In: Encyclopedia biostatistica. Boston: John Wiley & Sons; 1998.
[40] Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001;10:571–8.
[41] Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther 1993;73:386–95.
[42] Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol 2002;14:109–14.